STATISTICA, anno LXXV, n. 1, 2015

BAYESIAN HIERARCHICAL MODELS FOR MISALIGNED DATA: A SIMULATION STUDY

Giulia Roli [1]

Dipartimento di Scienze Statistiche, Università di Bologna, Bologna, Italia

Meri Raggi

Dipartimento di Scienze Statistiche, Università di Bologna, Bologna, Italia.

1. Introduction

Linking data collected at different scales, locations and dimensions constitutes a challenging topic of spatial analysis. The great interest on this area is mainly due to practical reasons, drawing on works from geographic, ecological, environmental, agricultural and geological fields. Indeed, the increasing availability of geographically referenced data calls for an exploitation of this information that avoids the implementation of new and expensive data collection. In particular, we refer to the case of European projects that often require ad hoc data collection for socio-economic or environmental analysis. Real studies usually consider areas defined more on the basis of problem features than on administrative boundaries. Significant time and money is required to collect original data, in many case limited by both the areas of interest and by very specific topics. This implies that the use of secondary data (for example administrative or census data) as covariates would be desirable and sometimes necessary.

Statistical literature generally refers to this topic as the overall problem of spatial misalignment or "incompatible" spatial data, meaning the analysis of data at a different level of spatial resolution with respect to the one originally collected. This concept includes an array of different problems with varying characteristics and numerous analytical solutions have been proposed to address them in the literature. For instance, the Modifiable Areal Unit Problem (MAUP) concerns the investigation of a variable's distribution at a new level of spatial aggregation; for data modeled through a spatial process, if the aim is to envision block averaging at different spatial scales, then we refer to the Change Of Support Problem (COSP). The ecological fallacy investigates the fact that relationships observed between variables measured at the aggregate level may not accurately reflect the relationship between these same variables measured at the individual level (Lawson, 2009; Banerjee *et al.*, 2004).

[1] Corresponding Author. E-mail: g.roli@unibo.it

This problem can occur in several combinations and consequently different solutions have been proposed in literature. In particular, spatial misalignment arises because we observe data referred to areas (or points) but the nature of the process is not coherent to them (for an exhaustive presentation of cases considering multidisciplinary studies, see Gotway and Young (2002)). Verdin *et al.* (2015) has followed a Bayesian kriging approach to address a point-to-area misalignment problem dealing with blending precipitation gauge data and satellite-derived precipitation estimates. Similar empirical problems have been dealt with by Sinclair and Pegram (2005) using conditional merging methods.

A different solution has been proposed by Lopiano *et al.* (2014) who considered a pseudo-penalized quasi-likelihood algorithm for kriging to align datasets in both point-to-point and point-to-area misalignment problems when the response variable is non-normally distributed. Other works focused on errors caused by spatial misalignment and corresponding measurements to properly estimate the relationship among variables that are misaligned in space ((Lopiano *et al.*, 2014; Szpiro *et al.*, 2011)). Similarly, Gryparis *et al.* (2008) faced this problem in epidemiologic research when the misaligned information refers to an environmental exposure with respect to an outcome, which is usually a human health response, measured at different spatial levels. Peng and Bell (2010) considered the problem of error measurement in time series analysis.

We focus our attention on a particular aspect of the more general topic of spatial misalignment, namely the area-to-area misalignment problem. This is the case of available data (typically counts or rates from administrative sources) which refer to spatial areas that are different from the ones of interest. The main aim is to convert the source information into target zones in order to avoid ad hoc data collections and then employ the imputed data in the subsequent analysis. The two misaligned spatial grids define regions that are too large to be considered as marked points and, thus, the methods previously described for point-to-area or point-to-point misalignments do not completely address the problem. The association of measurements observed in misaligned regions requires predicting the values of variables in regions in which they were not measured. This process, which is similar to kriging, is known as *areal interpolation* and has been particularly explored in the geographical literature (see, e.g., Goodchild *et al.* (1993)).

In the statistical literature, interesting approaches to areal interpolation have been proposed by Mugglin and Carlin (1998) and Mugglin *et al.* (1999). They present a hierarchical Bayesian method for interpolation, estimation and spatial smoothing of Poisson responses by exploiting information on a set of covariates on both grids. In this paper, we consider a similar, fully model-based, method proposed by Mugglin *et al.* (2000) over area-to-area misaligned grids which can be either nested or non-nested. The main advantage of this kind of approach is full inference (e.g. enabling interval estimates) for the distributions of target zone data. Here, we are interested in assessing model performance in the case of nested data grids and robustness towards model misspecifications. With this purpose, we generate artificial data inspired by a real study, on which we will apply the method for future development. We also provide a comparison of the results under different simulated scenarios.

The remainder of the paper is organized as follows. In Section 2, we introduce the practical reasons which motivate the need for incompatible spatial data integration through a case study description. Section 3 resumes the modeling framework proposed by Mugglin *et al.* (2000) and describes our assumptions. The simulation exercise based on the motivating example is presented in Section 4 and the corresponding results are provided in Section 5. Section 6 concludes with a discussion of the main results and future developments.

2. Motivating example

The concept of this paper arises from an analysis carried out in a study at the European level in the CLAIM project, funded by 7FP, (www.claimproject.eu). It sought to provide the knowledge base to support an effective Common Agricultural Policy (CAP) design to improve landscape management and, in particular, to offer insights into the ability of landscapes to contribute to the production of added value for society in rural areas. This objective is realized by investigating the connection between landscape features, rural economies and social characteristics of rural areas by means of nine case studies in different European countries.

In Italy, the case study refers to an area of ten municipalities in the province of Ferrara where part of the area is within the Po Delta Park. The need for statistical methods aimed at integrating misaligned data arises from the fact that the area of interest (Po Delta Park) does not coincide with the entire area of the municipalities in question, but is included within it. Hence we have two separate areas: the one of interest (the park) and one that includes ten municipalities and that is larger than the target one. Specific data were collected through an ad hoc survey and the sample unit was randomly selected in the larger area since only a population list at the municipality level was available. Figure 1 graphically summarizes the case study, where green, red and blue areas identify different zones of the Po Delta Park and where continuous lines represent the municipality boundaries.

In practice, the problem at hand relates to the use of data from both an ad hoc study and/or from an administrative source rather than the more accurate use of data related to the real target area. On one hand, data in the first category are readily available and geographically referenced. Furthermore, the secondary administrative data could be very useful as explanatory variables, but refer to different and larger areas than the selected case study sites. On another hand, secondary socio-economic data are not collected at the target area level and lists for selecting random sample units are not available. Possible examples of available administrative data are the number of organic farms and/or the number of financial subsidies received, whilst examples of covariates at the Po Delta grid level could be the number of farms and the number of young farmers.

3. Modeling framework

Let us consider two misaligned spatial grids, S_B and S_C, where the first one defines the so-called *source zones*, i.e. areas from which data of interest are available, and the second grid partitions regions for which data are to be imputed, namely

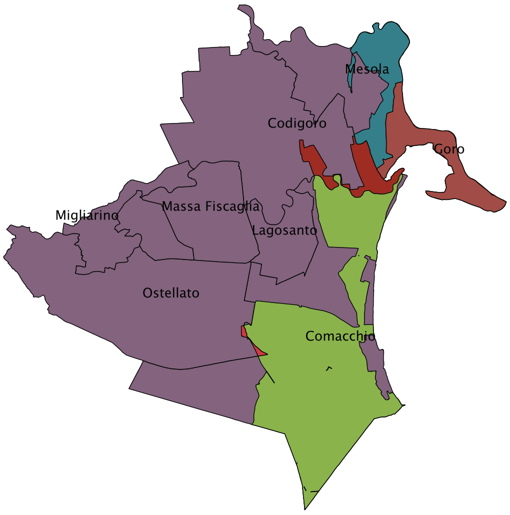

Figure 1 – Po Delta Park areas included in CLAIM project (green, red and blue zones identify different park subareas).

target zones. Thus, the first grid is often referred to as the *response grid* and the second one as the *explanatory grid*. We consider a special case with respect to the more general framework proposed by Mugglin *et al.* (2000), where one data grid contains the other, that is, following the motivating example, the administrative data coverage is higher than the whole area of interest, say $S_C \subset S_B$. In this case, C cells with portions lying outside of S_B will not exist.

In the first grid, regions are indexed by i (with $i = 1, \dots, I$) and denoted by B_i; similarly, for the second grid we have regions C_j, with $j = 1, \dots, J$. The intersection of the two grids creates *atoms*, i.e., cells identifying partitions of both the B_i and C_j regions. Atoms can be referenced relative to an appropriate B cell and denoted by B_{ik} (with $k = 1, \dots, K_i$) or, equally, to an appropriate C cell by C_{jl} (with $l = 1, \dots, L_j$). Thus, we can formally define the function f such that $f(B_{ik}) = C_{jl}$ and the inverse function g such that $g(C_{jl}) = B_{ik}$. Atoms with no intersection across the grids are *edge atoms*. In our case, only B-edge atoms can exist, say B_{iE}.

For each B_i source zone, we can observe the response Y_i. Referring to the notation introduced above, the main aim is then to convert Y_i to Y'_j, i.e. imputing values of Y to the target zones, by exploiting covariates that can be observed on the explanatory grid, X_j, and/or to the response grid, W_i. Let us consider only the availability of an X covariate, which is assumed to be an aggregated measurement similar to the response Y. Under this hypothesis the observed values of Y_i can be regarded as $\sum_k Y_{ik}$, where Y_{ik} are latent values for the atoms associated with B_i. Similarly, $X_j = \sum_j X_{jl}$, where X_{jl} are unobserved according to the atoms associated with C_j.

In order to specify the model, we assume Poisson distributions for the observed measurements

$$X_j \sim Poi(e^{\omega_j}|C_j|) \tag{1}$$

$$Y_i \sim Poi\left(e^{\mu_i} \sum_k |B_{ik}| h(X'_{ik}/|B_{ik}|; \theta_{ik})\right) \tag{2}$$

where $|A|$ denotes the area of a generic region A, ω_j and μ_i are random effects capturing spatial associations among the X_j's and the Y_i's, respectively, $h(\bullet)$ is a selected parametric function depending on θ_{ik} and adjusting an expected proportional-to-area allocation according to X'_{ik}, which are the values of X_{jl} associated to the response grid. As a result, the conditional distribution of the latent variables X_{jl} and Y_{ik} given the observed is a product multinomial

$$\left(X_{j1}, X_{j2}, \dots, X_{jL_j}\right) \sim mult(X_j; q_{j1}, q_{j2}, \dots, q_{jL_j}) \tag{3}$$

$$(Y_{i1}, X_{i2}, \dots, X_{iK_i}) \sim mult(Y_i; p_{i1}, p_{i2}, \dots, p_{iK_i}) \tag{4}$$

where $q_{jl} = \frac{|C_{jl}|}{|C_j|}$ and $p_{ik} = \frac{|B_{ik}| h(X'_{ik}/|B_{ik}|; \theta_{ik})}{\sum_k |B_{ik}| h(X'_{ik}/|B_{ik}|; \theta_{ik})}$.

For B-edge atoms, B_{iE}, there is no corresponding C_{jl}, thus a latent X'_{iE} is introduced the distribution of which is defined by the adjacent non-edge atoms

$$X'_{iE} \sim Poi(e^{\omega_i^*}|B_{iE}|) \tag{5}$$

where ω_i^* are additional spatial random effects to be associated to the others ω_j.

In a fully Bayesian setting, prior distributions for each parameter need to be specified. To capture the spatial nature of B_i, a Markov random field prior for the μ_i's can be adopted (Bernardinelli and Montomoli, 1992). In particular, if the adjacency form is considered, we obtain a conditional autoregressive (CAR) prior (Besag, 1974)

$$f\left(\mu_i|\mu_{i',i'\neq i}\right) \sim N\left(\sum_{i'} u_{ii'}\mu_{i'}/u_{i\cdot}, 1/(\lambda_\mu u_{i\cdot})\right) \tag{6}$$

where $u_{ii} = 0$, $u_{ii'} = u_{i'i}$, $u_{i\cdot} = \sum_{i'} u_{ii'}$ and $u_{ii'}$ equals 1 if $B_{i'}$ is a neighbor of B_i and 0 otherwise.

In the simplest form, we refer to an exchangeable prior that captures heterogeneity but not local clustering across areas, i.e. all $u_{ii'} = 1$ in equation 6. Similar considerations can be referred to in the set of spatial random effects for C_j regions, namely $\{\omega_j\omega_i^*\}$

$$f\left(\omega_j|\omega_{j',j'\neq j}\right) \sim N\left(\sum_{j'} v_{jj'}\omega_{j'}/v_{j\cdot}, 1/(\lambda_\omega v_{j\cdot})\right) \tag{7}$$

where $v_{jj} = 0$, $v_{jj'} = v_{j'j}$, $v_{j\cdot} = \sum_{j'} v_{jj'}$ and $v_{jj'}$ equals 1 if $C_{j'}$ is a neighbour of C_j and 0 otherwise or all $v_{jj'} = 1$ for exchangeable prior cases.

For other (hyper-) parameters, proper and vague priors are generally adopted (see Mugglin *et al.*, 2000, for more details). In particular, for λ_μ and λ_ω parameters proper gamma priors are usually adopted.

4. Simulation study

In order to test the method described above, we generate an artificial data set based on the practical purposes of the motivating example. As a first attempt, we consider a restricted number of areas for both grids, with the potential of being extended by using data from Global Positioning Systems (GPSs) and Geographical Information Systems (GISs). We select 3 source zones B_i and 4 target zones C_j, with $S_C \subset S_B$. The intersection of the two grids generates 3 B-edge atoms, B_{1E}, B_{2E} and B_{3E} and a total of 9 non-edge atoms, B_{ik} (with $k = 1, \ldots, K_i$ and $K_i = 3, 5, 4$) or C_{jl} (with $l = 1, \ldots, L_j$ and $L_j = 3, 2, 2, 2$). Figure 2 graphically resumes the grids.

Areas of regions and atoms and related variables X and Y are randomly generated, under two scenarios:

Scenario 1: both the spatial random effects follow exchangeable priors, i.e.

$$\mu_i \sim^{iid} N(\eta_\mu, \tau_\mu)$$
$$\{\omega_j, \omega{*}_i\} \sim^{iid} N(\eta_\omega, \tau_\omega) \tag{8}$$

Scenario 2: exchangeable prior for μ_i and a CAR distribution for $\{\omega_j, \omega_i^*\}$, with a correlation between adjacent areas $\rho = 0.9$, are fixed.

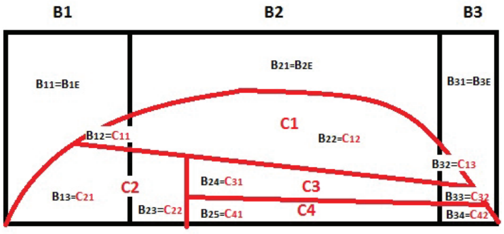

Figure 2 – Areal Data Misalignment: graphical representation of the problem considered in the simulation exercise (source zones in black contour, target areas in red).

TABLE 1
Values of the parameters for the generation of artificial data

Parameters	Values
η_μ	1.1
τ_μ	0.5
η_ω (Scenario 1)	4
τ_ω (Scenario 1)	1.2
η_ω (Scenario 2)	0
τ_ω (Scenario 2)	1
ρ (Scenario 2)	0.9
θ	1
$\lvert C_{jl}\rvert$	random from 100 to 500
$\lvert B_{iE}\rvert$	random from 100 to 500

The values of parameters assigned to generate data under the two scenarios, and according to the probability distributions described above, are summarised in Table 1. In particular, they are used to simulate data for the covariate X and the response Y for each atom through the multinomial distribution, given the parameters fixed for the distribution of $\{\mu_i\}$ and $\{\omega_j\omega_i^*\}$.

Parameter estimation is then implemented by considering the model specification introduced in Section 3, adopting a MCMC approach and using the software WinBUGS (Spiegelhalter *et al.*, 2003). Referring to the seminal work of Mugglin *et al.* (2000), we choose a function h for the computation of the p_{ik} values ensuring non-null estimates of the response Y on the atoms, i.e. $h(X'_{ik}/|B_{ik}|;\theta_{ik}) = X'_{ik}/|B_{ik}|+\theta_{ik}$ with $\theta_{ik} = \frac{\theta}{K_i|B_{ik}|}$ and $\theta = 1$. In order to assess model performance and robustness towards misspecifications, parameter estimations are carried out following this scheme, separately by the two following scenarios:

TABLE 2
Results - Scenario 1.

Parameters	True	Estimates	95% CI	% relative bias	coverages of 95% CI
Y'_1	27650	27590	27320 - 27820	-0.22	yes
Y'_2	2451	2426	2367 - 2492	-1.02	yes
Y'_3	8246	8263	8192 - 8352	0.21	yes
Y'_4	53793	53690	53190 - 54280	-0.19	yes
$Y'_{1,1}$	2447	2375	2309 - 2441	-2.94	no
$Y'_{1,2}$	13303	13160	12800 - 13470	-1.07	yes
$Y'_{1,3}$	11900	12050	11850 - 12280	1.26	yes
$Y'_{2,1}$	75	73	69 - 77	-2.67	yes
$Y'_{2,2}$	2377	2353	2295 - 2418	-1.01	yes
$Y'_{3,1}$	5014	4969	4836 - 5102	-0.90	yes
$Y'_{3,2}$	3233	3294	3227 - 3365	1.89	yes
$Y'_{4,1}$	34415	33960	33200 - 34750	-1.32	yes
$Y'_{4,2}$	19378	19730	19380 - 20050	1.82	no

Scenario 1:

- estimation with assumptions of data generation.

Scenario 2:

- estimation with assumptions of data generation;
- estimation wrongly assuming both of the spatial random effects follow exchangeable priors.

5. Results

The results of the simulation exercise are reported with respect to the parameters of interest in response Y imputed for both target zones, Y'_j, and non-edge atoms associated to the explanatory grid, Y'_{jl}.

As far as Scenario 1 is concerned, in Table 2 we show and compare true (ϕ) and estimated ($\widehat{\phi}$) parameters, together with 95% credibility intervals (CI), percentage relative bias ($\frac{\phi-\widehat{\phi}}{\phi} * 100$) and coverages of 95% CI (yes or no).

The estimates seem to be quite precise, with acceptable relative biases (varying from a minimum of -2.94% to a maximum of 1.89%). Only two cases of non-coverage of 95% Credibility Intervals are yielded.

For Scenario 2, we mainly compare the results obtained through the two assumptions concerning the spatial random effects μ_i and $\{\omega_j, \omega_i^*\}$. Estimates are reported in Table 2 together with 95% Credibility Intervals (CI), minimum and maximum values of percentage relative bias and the Deviance values in order to test and compare the goodness-of-fit, separately by the two models.

By comparing the two model specifications, it is noteworthy that the true model, i.e. considering a CAR distribution for $\{\omega_j, \omega_i^*\}$ parameters, is properly identified as the best one through the Deviance criterion. Moreover, corresponding relative biases are narrower than those yielded by the other model specifications.

TABLE 3
Results - Scenario 2.

		CAR prior		**Exchangeable prior**	
Parameters	True	Estimates	(95% CI)	Estimates	95% CI
Y_1'	32544	31058	27774 - 33386	29340	24170 - 33890
Y_2'	24372	24138	21730 - 26760	28180	21990 - 33850
Y_3'	30701	29480	26940 - 31232	29720	24140 - 34350
Y_4'	51190	48200	44450 - 49790	49370	39810 - 57110
$Y_{1,1}'$	2869	3088	2431 - 3891	2249	1591 - 2946
$Y_{1,2}'$	15723	15420	13400 - 17470	18080	14070 - 21960
$Y_{1,3}'$	13952	12550	10930 - 13900	9011	7236 - 10900
$Y_{2,1}'$	748	808	625 - 1021	587	407 - 785
$Y_{2,2}'$	23624	23330	20810 - 25940	27590	21490 - 33120
$Y_{3,1}'$	18842	18520	16460 - 20900	21860	17020 - 26340
$Y_{3,2}'$	11859	10960	9685 - 11970	7865	6341 - 9507
$Y_{4,1}'$	33543	31770	28450 - 35390	37580	29220 - 45130
$Y_{4,2}'$	17647	16430	14310 - 17740	11790	9568 - 14070
% relative bias - min	-10.05			-35.41	
% relative bias - max	7.95			16.79	
deviance	79.74			80.04	

The estimates obtained by misspecificating the distribution of spatial random effects, i.e. by wrongly assuming that both the spatial random effects follow exchangeable priors, do not seem to be particularly biased and the 95% Credibility Intervals all cover the corresponding true parameter values. As a result, we can conclude that the method is quite robust for this kind of model misspecification.

Finally, the comparison of the results under the two scenarios shows that the accuracy of the estimates decreases when the complexity of the data (and, consequently of the model) increases. Indeed, the average absolute relative bias amounts to 1.27, in the simplest case where both the spatial random effects follow exchangeable priors (Scenario 1); then, it grows into 5.05 and 18.27, when a spatial structure of random effects is considered (Scenario 2).

6. Final remarks

In this work, we considered the problem of combining information from different data sources focusing on spatially misaligned data. A hierarchical Bayesian model is used to convert the source information to target zones by exploiting a set of covariates on both grids. We applied this method to simplify simulated data generated to resemble a real study. In particular, we referred to the case where the administrative data grid contains the whole areal grid of interest.

The estimates we obtained appear to be quite precise, with acceptable relative biases. Moreover, the method we used seems to be quite robust for misspecifications with regard to the distribution of the spatial random effects. Thus, it is appropriate to apply this method of areal interpolation to the real data that is the inspiration for our research question. Moreover, the fully model-based approach enables us to adopt an inferential, and not only descriptive, perspective of the results (Mugglin and Carlin (1998); Mugglin *et al.* (1999, 2000)).

For future development of this research question, we would assess method robustness by considering a larger number of areas in both grids and the effects of model misspecifications. In particular, we would focus on the impact on model performance when other assumptions are imposed (e.g. on both the priors and the observed measurements). Finally, we would test the effect of different values of correlation between adjacent areas, considering CAR distributions for both spatial random effects and comparing different specifications of function h, i.e. differently adjusting the expected proportional-to-area allocation.

References

S. Banerjee, B. P. Carlin, A. E. Gelfand (2004). *Hierarchical Modeling and Analysis for Spatial Data.* Chapman and Hall, London.

L. Bernardinelli, C. Montomoli (1992). *Empirical bayes versus fully bayesian analysis of geographical variation in disease risk.* Statistics in Medicine, 11.

J. Besag (1974). *Spatial interaction and the statistical analysis of lattice systems (with discussion).* Journal of the Royal Statistical Society, series B, 36.

M. F. Goodchild, L. Ansellin, U. Deichmann (1993). *A framework for the areal interpolation of socioeconomic data.* Environment and Planning, A, no. 25, pp. 383–387.

C. A. Gotway, L. J. Young (2002). *Combining incompatible spatial data.* Journal of the American Statistical Association, 97, no. 458, pp. 632–648.

A. Gryparis, C. J. Paciorek, A. Schwartz, B. Coull (2008). *Measurement error caused by spatial misalignment in environmental epidemiology.* Biostatistics, 10.

A. B. Lawson (2009). *Bayesian disease mapping.* CRC press.

K. Lopiano, L. Young, C. Gotway (2014). *A pseudo-penalized quasi-likelihood approach to the spatial misalignment problem with non-normal data.* Biometrics, 70.

A. Mugglin, B. Carlin (1998). *Hierarchcal modeling in geographic information systems: Population interpolation over incompatible zones.* Journal of Agricultural, Biological, and Environmental Statistics, 3.

A. Mugglin, B. Carlin, A. Gelfand (2000). *Fully model-based approaches for spatially misaligned data.* Journal of the American Statistical Association, 95, no. 451, pp. 877–887.

A. Mugglin, B. Carlin, L. Zhu, E. Colon (1999). *Bayesian areal interpolation, estimation, and smoothing: An inferential approach for geographic information systems.* Environment and Planning A, 3, no. 1, pp. 1337–1352.

R. Peng, L. Bell (2010). *Spatial misalignment in time series studies of air pollution and health data.* Biostatistics, 11, no. 4, p. 720740.

S. Sinclair, G. Pegram (2005). *Combining radar and rain gauge rainfall estimates using conditional merging.* Atmospheric Science Letters, 6, no. 1, pp. 19–22.

D. Spiegelhalter, A. Thomas, N. Best, D. Lunn (2003). *WinBUGS User Manual, Version 1.4.*

A. Szpiro, L. Sheppard, T. Lumley (2011). *Efficient measurement error correction with spatially misaligned data.* Biostatistics, 12, no. 4, p. 610623.

A. Verdin, B. Rajagopalan, W. Kleiber, C. Funk (2015). *A bayesian kriging approach for blending satellite and ground precipitation observations.* Water Resources Research, 51.

Summary

In this paper, the problem of combining information from different data sources is considered. We focus our attention on spatially misaligned data, where available information (typically counts or rates from administrative sources) refers to spatial units that are different from the ones of interest. A hierarchical Bayesian perspective is considered, as proposed by Mugglin *et al.* in 2000, to provide a fully model-based approach in an inferential, and not only descriptive, sense. In particular, explanatory covariates are arranged to be modeled according to spatial correlations through a conditionally autoregressive prior structure. In order to assess model performance and its robustness we generate artificial data inspired by a real study and a simulation exercise is then carried out.

Keywords: Bayesian analysis; Misaligned data; Linking spatial information.